Machine Translation from an Intercomprehension Perspective

Yu Chen **Tania Avgustinova**
Department of Language Science and Technology
Saarland University, Saarbrücken, Germany
{yuchen, tania}@coli.uni-sb.de

Abstract

Within the first shared task on machine translation between similar languages, we present our first attempts on Czech to Polish machine translation from an intercomprehension perspective. We propose methods based on the mutual intelligibility of the two languages, taking advantage of their orthographic and phonological similarity, in the hope to improve over our baselines. The translation results are evaluated using BLEU. On this metric, none of our proposals could outperform the baselines on the final test set. The current setups are rather preliminary, and there are several potential improvements we can try in the future.

1 Introduction

A special type of semi-communication can be experienced by speakers of similar languages, where all participant use their own native languages and still successfully understand each other. On the other hand, in countries with more than one official languages, even if these languages are mutually intelligible. While it is a common practice to use English as a pivot language for building machine translation systems for under-resourced language pairs. If English turns out to be typologically quite distant from both the source and the target languages, this circumstance easily results in accumulation of errors. Hence, the interesting research question is how to put the similarity between languages into use for translation purposes in order to alleviate the problem caused by the lack of data or limited bilingual resources.

Slavic languages are well-known for their close relatedness, which may be traced to common ancestral forms both in the oral tradition and in written text communication. Sharing many common features, including an inventory of cognate sound-meaning pairings, they are to various degrees mutually intelligible, being at the same time so different that translating between them is never an easy task. For example, all Slavic languages have rich morphology, but inflections systematically differ from one language to another, which makes it impossible to have a uniform solution for translating between them or to a third language.

We chose to work on the language pair Czech-Polish from the West Slavic subgroup. In an intercomprehension scenario, when participants in a multilingual communication speak their native languages, Czechs and Poles are able to understand each other to a considerable extent, mainly due to objectively recognisable and subjectively perceived linguistic similarities. As Czech-English and Polish-English translation pairs are challenging enough, this naturally motivates the search for direct translation solutions instead of a pivot setup.

We first briefly introduce the phenomenon of intercomprehension between Slavic languages and our idea how to take advantage of it for machine translation purposes. The next section spreads out our plans on Czech-Polish translation by exploring the similarities and differences between the two languages. Then, we explain how we organized the experiments that lead to our submissions to the shared task. We conclude with a discussion of the translation results and an outlook.

2 Slavic Intercomprehension

Intercomprehension is a special form of multilingual communication involving receptive skills when reconstructing the meaning in inter-lingual contexts under concrete communicative situation It is common practice for millions of speakers, especially those of related languages. In order to interpret the message encoded in a foreign but related language, they rely on linguistic and non-linguistic elements existing for similar situations

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 192–196
Florence, Italy, August 1-2, 2019.

in their own linguistic repertoire.

Languages from the same family exhibit systematic degrees of mutual intelligibility which may be in many cases asymmetric. Czech and Polish belong to the West Slavic subgroup and related both genetically and typologically. It could be shown, for example, that the Poles understood written Czech better than the Czechs understood written Polish, while the Czechs understood spoken Polish better than the Poles understood spoken Czech (Golubović, 2016). How can this be useful for machine translation? In order to tackle the Czech-to-Polish machine translation problem from an intercomprehension point of view, we currently focus on orthographic and phonological similarities between the two languages that could provide us with relevant correspondences in order to establish inter-lingual transparency and reveal cognate units and structures.

3 Approach

3.1 Orthographic correspondences

Both orthographic systems are based on the Latin alphabet with diacritics, but the diacritical signs in the two languages are rather different. Czech has a larger set of letters with diacritics, while Polish uses digraphs more often. There are two basic diacritical signs in Czech: the acute accent (´) used to mark a long vowel and the háček (ˇ) in the consonants which becomes the acute accent for d’ and t’. The diacritics used in the Polish alphabet are the kreska (graphically similar to the acute accent) in the letters ć, ń, ó, ś, ź; the kreska ukośna (stroke) in the letter ł. ; the kropka (overdot) in the letter ż; and the ogonek ("little tail") in the letters ą, ę. The Czech letters á, č, d’, é, ě, ch, í, ň, ř, š, t’, ú, ů, ý, ž as well as q, v, and x do not exist in Polish, and the Polish letters ą, ć, ę, ł, ń, ś, w, ż and ź are not part of the Czech alphabet.

In a reading intercomprehension scenario, it is natural for people to simply ignore unknown elements around graphemes that they are familiar with. That is, when facing unknown alphabet with “foreign” diacritical signs, the reader is most likely to drop them and treat the respective letters as the corresponding plain Latin ones. Experiments showed that efficiency of intercomprehension is significantly improved if the text is manually transformed to mimic the spelling in the reader’s language (Jágrová, 2016). However, such rewriting requires a huge effort from a bilingual linguist and cannot be easily applied to large amount of data. An alternative to the manual rewriting is to utilize the correspondence rules using Minimum Description Length (MDL) principle (Grünwald, 2007). Most of the around 3000 rules generated from a parallel cognate list of around 1000 words are not deterministic. We use only the rules converting Czech letters that do not exist in Polish, as listed in Table 1, to avoid over-transformation.

CZ	PL
Áá	Aa
Čč	Cz cz
Ďď’	Dź dź
Ěě	Je je
Éé	Ee
Íí	Ii
Ňň	Nn
Řř	Rz rz
Šš	Sz sz
Ťť’	Ćć
Ůů	Óó
Vv	Ww
Xx	Ks ks
Ýý	Yy
Žž	Źź

Table 1: Orthographic correspondence list

3.2 Phoneme correspondences

Czech and Polish are both primarily phonemic with regard to their writing system, which is reflected in the alphabets. That is, graphemes consistently correspond to phonemes of the language, but the relation between spelling and pronunciation is more complex than a one-to-one correspondence. In addition, Polish uses more digraphs, such as ch, cz, dz, dź, dż, rz, and sz. In both languages, some graphemes have been merged due to historical reasons and at the same time some changes in phonology have not been reflected in spelling.

It is well-known, that people often try to pronounce the foreign texts in a way closer to their own language. Moreover, hearing the correct pronunciation sometimes helps them to infer the meaning more easily, in particular, loanwords / internationalisms and the pan-Slavic vocabulary.

To be able to make use of phonological information within a machine translation system, we

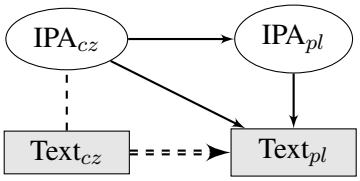

Figure 1: Phoneme-infused translation setup

propose a multi-source multi-target structure as shown in Figure 1, which considers the IPA transcription of the text as a text in a "new" closely related language. More specifically, for the translation from Czech to Polish, the source languages of the multilingual system include ordinary Czech and IPA-transcribed Czech. So, three different translation paths are competing with each other to produce the final translations.

4 Experiments

4.1 Data and baselines

We used the provided parallel corpora Europarl, Wiki Titles, JRC-Acquis and the monolingual corpus News Crawl 2018 for Polish. We extract randomly two disjoint subsets from the development set of the shared task: one with 2000 sentences and another one with 1000 sentences. During the development phase, all systems are optimized for the BLEU score on the first set and the second set is used as a blind test set. The results reported in the next section refer to BLEU scores (Papineni et al., 2002) on the official test set unless specified otherwise.

For the purpose of more thorough comparisons, we build three baseline systems in different paradigms, one phrase based statistical machine translation system (PBSMT) with the Moses toolkit (Koehn et al., 2007) and two neural machine translation (NMT) system with the marian toolkit (Junczys-Dowmunt et al., 2018). All baselines apply the same pre- and post-processing steps. Preprocessing consists of tokenization, truecasing and removing sentences with more than 100 tokens. Postprocessing consists of detruecasing and detokenization. All these steps use scripts included in the Moses toolkit.

The PBSMT baseline uses both the target side of the parallel corpora and the monolingual corpus provided for the language model. 5-gram language models are first built individually from each corpus and then interpolated with KenLM (Heafield et al., 2013) given the development. We run *fast_align* (Dyer et al., 2013) on the parallel corpora to obtain word alignments in both directions. Then, phrase pairs with less than 6 tokens are extracted to construct a translation model based on the alignments. Weights for the features in the translation model are determined with the Minimal Error Rate Training (MERT) (Och, 2003).

A byte pair encoding (BPE) (Sennrich et al., 2015) is applied to the training data to reduce the vocabulary to 36,000 units for the NMT systems. The first NMT system utilized only the parallel data. It is a single sequence-to-sequence model with single-layer RNNs in both the encoder and the decoder. The embedding size is 512 and the RNN state size is 1024.

The architecture of our second NMT baseline follows the architecture described in (Vaswani et al., 2017). We first train a shallow model from Polish to Czech with only the parallel corpora in order to translate the complete monolingual Polish corpus into Czech for a synthesized parallel corpus, which is concatenated with the original data to produce new training data (Sennrich et al., 2016). We then train four left-to-right (L2R) deep Transformer-based models and four right-to-left (R2L) models. The ensemble decoder combines the four L2R models to generate an n-best list, which is rescored using the R2L models to produce the final translation.

System	BLEU
PBSMT	11.58
Deep RNN	9.56
Transformer-based	
+ Ensemble	
+ Rerenking	13.46

Table 2: Czech-Polish baselines on development test

Table 2 lists the BLEU scores of the baselines. To our surprise, the simple “old-fashioned” PBSMT system surpassed the RNN-based NMT system and was close to the Transformer-based ensemble. In fact, the translations produced by the Transformer-based NMT are not significantly better than those from the PBSMT.

4.2 Translation results

The outcome of various experiments based on the produced baseline systems is presented here by first looking into the PBSMT and then into the

Transformer-based NMT.

Note that we actually made a mistake inserting the source segments into each translation segment for the final submission. Therefore, the results reported here are all produced after the submission by re-evaluating the clean sgm files. All the scores are cased BLEU scores calculated with the nist evaluation script *mteval-v14.perl*.

4.2.1 Modifying PBSMT

Our PBSMT experiments start with applying a joint BPE model to the training sets, both parallel and monolingual, similarly to the approach introduced by (Kunchukuttan and Bhattacharyya, 2016).

Given the lexical similarity between Czech and Polish, a joint BPE model identifies a small cognate vocabulary of subwords from which words in both languages can be composed. This step eventually identifies the orthographic correspondences as described in Section 3.1. BPE operations expand the sentences (ratio shown in Table 3), therefore we increase the order of the language model from 5 to 7 and the maximal phrase length in the translation model from 5 to 6. We also apply character replacements following the list shown in Table 1. Table 4 lists the results of the 3 combinations of these two operations. The translation does not seem to benefit from the character replacement. The BPE operation does not improve the system over the test set either, despite that a minor change was recorded on the development test set.

Corpus	%
Acquis	119.23
Europarl	118.09
WikiTitles	238.96
News 2018	145.81

Table 3: Sentence expansion due to BPE operatios
Sentence length ratio (%)

	BLEU$_{dev}$	**BLEU**$_{test}$
PBSMT baseline	11.58	9.62
+ BPE	12.21	7.90
+ replacement	11.53	5.31*
+ BPE + replacement	11.89	6.71*

Table 4: Translation results of PBSMT systems
* marks the system trained on partial development set

	BLEU$_{dev}$	**BLEU**$_{test}$
Transformer baseline	13.46	11.54
+replacement	13.33	11.25*
Phoneme-based		4.90
+reranking		5.88

Table 5: Translation results of NMT systems
* marks the system trained on partial development set

4.2.2 Modifying NMT

Table 5 shows the results from the second group of experiments. We have applied the same character replacement to our Transformer-based NMT system, but the impact is again minimal.

As for the phoneme-based system, we first convert all the data into IPA transcriptions using the finite state transducer (FST) model from (Deri and Knight, 2016) with the Carmel toolkit (Graehl, 2019) according to the languages. Consequently, we have 4 versions of the same messages: Czech texts, Czech IPA transcriptions, Polish texts and Polish IPA transcriptions. Considering proximity between the texts and the transcriptions, we use two separate BPE's: one for the texts and another one for the transcriptions. To construct the multiway NMT system illustrated in Figure 1, we gather 3 pairs of parallel texts together: (IPA$_{cs}$, Text$_{pl}$), (IPA$_{cs}$, IPA$_{pl}$) and (IPA$_{pl}$, Text$_{pl}$). We add tokens to each source sentences to mark the source and target sentence language (Ha et al., 2016). Then, such a concatenated parallel corpus is used to train a Transformer-based NMT system. The test set is sent through this multiway system to create an n-best list, which is scored with the original Transformer-based baseline.

Due to deadline constraints, we do not have enough time for thorough experiments on this setup. Such a design seems to degrade the system significantly, but it is also clear that such an architecture is producing very different predictions for the translation.

5 Discussions

This contribution describes our submission to the shared task on similar language translation. It is our first attempt to make use of orthographic and phonological correspondences between two closely related languages, Czech and Polish, inspired by their mutual intelligibility.

The current setups are rather preliminary. Currently, none of our methods improves the baselines

on the final test set. There are several potential improvements we can try in the future.

A fixed short replacement list we used is just a small portion of the orthographic correspondence rules. We are considering to integrate the orthographic correspondences with a BPE model as our next step.

Regarding the phoneme based system, the next thing to investigate is the choice of grapheme-to-phoneme (g2p) tools. It is not yet clear which g2p tool and which phoneme transcription set suit our purpose the best. Grouping similar phonemes is one of the potential direction to explore.

References

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 399–408.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Jelena Golubović. 2016. *Mutual intelligibility in the Slavic language area*. Ph.D. thesis, University of Groningen.

Jonathan Graehl. 2019. Carmel finite-state toolkit.

Peter D Grünwald. 2007. *The minimum description length principle*. The MIT Press.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Klára Jágrová. 2016. Adaptation towards a reader's language: The potential for increasing the intelligibility of polish texts for czech readers. In *12th European Conference on Formal Description of Slavic Languages*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Learning variable length units for smt between related languages via byte pair encoding. *arXiv preprint arXiv:1610.06510*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.